Short-range Forecasting Research

Scientific Note No. 24

ITERATIVE ANALYSIS USING COVARIANCE FUNCTIONS AND FILTERS

by

Andrew Lorenc

July 1991

ORGS UKMO S Duplicate

National Meteorological Library
FitzRoy Road, Exeter, Devon. EX1 3PB

Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

METEOROLOGICAL OFFICE
C15876
LIBRARY

Short-Range Forecasting Research Division
Scientific Note No. 24

ITERATIVE ANALYSIS USING COVARIANCE FUNCTIONS AND FILTERS

by

Andrew Lorenc

July 1991

Short-Range Forecasting Research Division
Meteorological Office
London Road
Bracknell
Berkshire RG12 2SZ
ENGLAND

N.B This paper has not been published. Permission to quote from it must be obtained from the Assistant Director of the above Met. Office branch.

Iterative Analysis using Covariance Functions and Filters

by Andrew Lorenc

Meteorological Office (S)

SUMMARY

An earlier review paper of analysis methods for Numerical Weather Prediction showed how many methods could be derived from the same basic analysis equation, expressing the Bayesian probability that any state is the true state. Prior probabilities depend on the error covariance of the background prior estimate, and these covariances figure in most solutions. This paper extends the analysis of iterative solution methods (one of which is known as the "Successive Correction" method). In a simple example the methods are compared with a direct solution (the "Optimal Interpolation" method); the iterative methods can be equally "optimal". If the number of iterations is limited, or the background weighting is omitted, then the methods normalizing the increments in observation space are shown to be more reliable than those normalizing in grid-point space (like the traditional successive correction method).

The covariances in these methods perform a filtering function; it is possible to replace them by a filter acting on the analysis increments. Iterative methods using such a filter are derived, corresponding exactly to the iterative methods using covariance functions. A simple and efficient recursive filter is described, and applied to the same example. The analysis using a two-pass filter is almost identical to that using a "Second Order Auto-Regressive" covariance function. A filter with many passes corresponds to a Gaussian shaped covariance function.

Approximate filters can be devised to model the effect of observational error correlations, with an accuracy probably as good as the knowledge of the real correlations. With this filter, the iterative methods can extended to deal effectively with data from remote sensing instruments.

Published iterative methods (many originally empirically derived) are reviewed, and fitted into the optimal theory. Practical applications are discussed.

1 INTRODUCTION

If, as is usually the case, there are insufficient observations to define a field which we wish to analyse, then additional information must be used. In practice we know, prior to knowing the observations, that some states are more likely than others; this can be expressed as a probability density function (p.d.f.). Lorenc (1986) showed how Bayes' Theorem could be used to derive an equation for the posterior p.d.f., taking into account the information added to the prior p.d.f. by the observations. Finding the state which maximizes the posterior p.d.f. is equivalent to finding that which minimizes a penalty function; from this various methods for objective analysis can be derived. In Numerical Weather Prediction, we always have a prior estimate of the field to be analysed, which we call the background field. Often, much of the rest of our prior knowledge concerns the smoothness of deviations from this field. This smoothness can be expressed in terms of the spatial covariance of the background errors. Optimal Interpolation (OI) uses a covariance function to set up, and solve directly, the minimization problem. Direct solution requires either very large matrix operations, or a severe restriction on the number of data used at once. There are several methods of iterative solution which avoid this; in section 2 we summarize the derivation of Lorenc (1986), and illustrate the properties of solution methods in a simple example.

Another way of expressing smoothness is by using filter operators. Particularly for dense observations and grid, this requires less computation. In section 3 we express the background error covariance function in terms of a smoothing operator. We analyse the same example using a recursive filter instead of a covariance function; the same selection of iterative methods can be used.

Remote sensing, where one instrument and retrieval method is used to make many observations, often provides observations with correlated errors. In section 4 we extend the examples to such a case. A method, using filters, is proposed for approximation the matrix inversion that arises in an exact solution.

Many of the methods presented in this paper are similar or identical to methods which have been applied in practice. In section 5 we review these, and discuss possible further applications.

2 ITERATIVE METHODS FOR OPTIMAL INTERPOLATION

For computation, we have to choose a discrete representation of the continuous field; $\mathbf{x}$ is one possible field in this representation. Knowledge that some states are less likely, implies some knowledge of the probability density function $P_b(\mathbf{x})$, that $\mathbf{x}$ is the representation of the true field $\mathbf{x}_t$. Often our prior knowledge can be expressed as a background field $\mathbf{x}_b$, which differs from $\mathbf{x}_t$ by errors which have a Gaussian distribution. Then

$$P_b(\mathbf{x}) \propto \exp(-\tfrac{1}{2}(\mathbf{x}_b-\mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x})). \tag{1}$$

B is the background error covariance matrix (assuming $\mathbf{x}_b$ is unbiased):

$$\mathbf{B} = \langle(\mathbf{x}_b-\mathbf{x}_t)(\mathbf{x}_b-\mathbf{x}_t)^T\rangle \tag{2}$$

In Lorenc (1986) the analysis equation is derived for a discrete set of observations $\mathbf{y}_o$, with instrumental error p.d.f. such that the probability of getting $\mathbf{y}_o=\mathbf{y}$ is given by

$$P_o(\mathbf{y}) = \exp(-\tfrac{1}{2}(\mathbf{y}-\mathbf{y}_t)^T\mathbf{O}^{-1}(\mathbf{y}-\mathbf{y}_t)). \tag{3}$$

K is a generalized interpolation operator from model space to observation space; because of errors of representativeness, this is not exact. Thus the p.d.f. that $\mathbf{y}=\mathbf{y}_t$ is

$$P_f(\mathbf{y}) = \exp(-\tfrac{1}{2}(\mathbf{y}-K(\mathbf{x}_t))^T\mathbf{F}^{-1}(\mathbf{y}-K(\mathbf{x}_t))). \tag{4}$$

The optimal analysis, $\mathbf{x}_a$, is that which maximizes the posterior probability found by combining these using Bayes' Theorem; for the Gaussian p.d.f. defined above this is equivalent to minimizing a penalty function J given by

$$J = \tfrac{1}{2}(\mathbf{x}_b-\mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) + \tfrac{1}{2}(\mathbf{y}_o-K(\mathbf{x}))^T(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}_o-K(\mathbf{x})). \tag{5}$$

A solution to this may be found by iterating

$$\mathbf{x}_{[u+1]} = \mathbf{x}_{[u]} - (J'')^{-1}J' \tag{6}$$

where

$$J' = -\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) - \mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}_o-K(\mathbf{x})), \tag{7}$$

$$J'' = \mathbf{B}^{-1} + \mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}\mathbf{K}, \tag{8}$$

and K is the tangent linear operator of K. Substituting gives :

$$\mathbf{x}_{[u+1]}=\mathbf{x}_{[u]}+(\mathbf{B}^{-1}+\mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}\mathbf{K})^{-1}(\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}_{[u]})+\mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}_o-K(\mathbf{x}_{[u]}))). \tag{9}$$

Moving a factor B, and defining weights W and normalization factor Q by

$$\mathbf{W} = \mathbf{B}\mathbf{K}^T(\mathbf{O}+\mathbf{F})^{-1} \tag{10}$$

$$\mathbf{Q} = (\mathbf{I}+\mathbf{W}\mathbf{K})^{-1} \tag{11}$$

gives:

$$x_{[u+1]}=x_{[u]}+Q(\ W(y_o-K(x_{[u]}))\ +\ (x_b-x_{[u]})\ ). \tag{12}$$

(Lorenc (1986) has a bracket mispositioned in (39)). Whatever the value of Q, if an equation of this form converges to a solution x_a, then at convergence

$$W(y_o-K(x_a))\ +\ (x_b-x_a)\ =\ 0. \tag{13}$$

This is the same solution as give directly by OI.

The analysis error covariance A is given by

$$A^{-1}\ =\ B^{-1}\ +\ K^T(O+F)^{-1}K. \tag{14}$$

This gives

$$A\ =\ Q\ B. \tag{15}$$

So Q can be thought of as expressing the relative reduction of error due to the observations.

To illustrate the various solution methods, as we derive them, we use a simple example of the analysis of a scalar field, given a square array of nine observations (figure 1). (It is important when judging these iterative methods to use a realistic two-dimensional distribution of observations; the similar one-dimension example converges much more quickly). We take $x_b=0$, and a background error correlation between points a distance r apart given by a Second Order Auto-Regressive (SOAR) function (Thiebaux *et al.* 1986, Balgovind *et al.* 1983):

$$\mu(r)\ =\ (1+r/s)\ \exp(-r/s). \tag{16}$$

The elements of B are covariances, given by $\varepsilon_b^2\mu$. Observation errors are assumed to be uncorrelated, with variance (ε_o^2) equal to 0.25 of the background error variance (ε_b^2). Rather than evaluating μ between grid-points, to get B, and then interpolation using K to get W, we do as in OI and evaluate (16) directly between observation and grid points to evaluate W.

The final net weight given to observation i at point k can be defined (even for an analysis method which does not use weights) as

$$n_{ki}\ =\ \frac{\partial\ x_k}{\partial y_i}. \tag{17}$$

If observation i were alone, it would have an optimal weight given by

$$n_{ki}\ =\ \mu_{ki}/(1+\varepsilon_o^2/\varepsilon_b^2) \tag{18}$$

Figure 2a shows this for the points along the line shown dashed in figure 1, and the observation points on the line. If we write the weights n_{ki} as a

matrix **N**, then we can write an expression for the error covariance of the resulting field

$$\mathbf{A} = \mathbf{B} - 2\,\mathbf{NKB} + \mathbf{N}(\mathbf{KBK}^T + (\mathbf{O}+\mathbf{F}))\mathbf{N}^T. \qquad (19)$$

This expression is valid whether or not the weights **N** are optimal. Figure 2b shows the optimal weights for the observations taken together, the sum of the weights for all nine observations, and the relative error reduction; the diagonal elements of (19) divided by the background error variance ε_b^2.

The form for Q given by (11) will give rapid convergence, but is expensive to compute. We can find an approximation that converges rapidly for a selected mode such as **m** by finding a diagonal Q such that

$$\mathbf{m} = \mathbf{Q}\,(\mathbf{I}+\mathbf{WK})\,\mathbf{m}. \qquad (20)$$

Convergence is guaranteed for all modes if all eigenvalues of (I-Q(I+WK)) have modulus less than one. For a simple scalar field, without the correlated observational errors or multivariate covariances discussed in sections 4 and 5, we will see in section 3 that analysis is equivalent to a low pass filter, damping small scales in the increments. The largest scales are undamped and will have the largest eigenvalues. So we can make **m** a constant field, giving a formula for the elements of the appropriate diagonal approximation

$$q_{kk} = \left(\sum_i w_{ki} + 1 \right)^{-1}, \qquad (21)$$

where w_{ki} is the element of W corresponding to the weight given to observation i at point k during each iteration; it is different from n_{ki}. q_{kk} is plotted in figure 2a. It can be compared with the relative error reduction in figure 2b. Because it treats all scales like large scales, q_{kk} overestimates the error reduction. Figure 3 shows the weights at various stages of the iteration of (12) using (21). Note that it is converging towards the optimal answer, but in early iterations the weights do not have the form shown in figure 2a. The effective weights are significantly sub-optimal in the area away from the observations, where extrapolation is needed. In this region the analysis is less accurate than the background; the relative analysis error shown in figure 3 is greater than one.

For a more general multivariate analysis, for instance of mass and wind, it may be that states that do not obey a relationship such as geostrophy are very unlikely. This will be reflected in the covariances B, and hence in W through (10). However the approximate normalization factor given by (21), illustrated in figure 2a, varies over the grid, so the properties of W need not carry through to the analysis increments. This is the reason why the intermediate

iteration weights in figure 3 are broader than optimal. We can avoid this by manipulating (9) so as to have a normalization in observation space $\tilde{Q}$

$$\tilde{Q} = (I+KW)^{-1}. \tag{22}$$

The iteration can then be written

$$\mathbf{x}_{[u+1]} = \mathbf{x}_{[u]} + W\tilde{Q}(\mathbf{y}_o - K(\mathbf{x}_{[u]})) + Q(\mathbf{x}_b - \mathbf{x}_{[u]}). \tag{23}$$

The elements of $\tilde{Q}$ can be approximated as in (21), with k being this time an observation position. However applying inconsistent approximations to Q and $\tilde{Q}$ causes (23) not to converge to $\mathbf{x}_a$. We need to convert the background term to observation space, to be normalized by $\tilde{Q}$ also; we do this by adding it into iterated estimates of $\mathbf{y}_t$:

$$\begin{aligned} \mathbf{y}_{[u+1]} &= \mathbf{y}_{[u]} + \tilde{Q}(\mathbf{y}_{[u]} - K(\mathbf{x}_{[u]})) \\ \mathbf{x}_{[u+1]} &= \mathbf{x}_{[u]} + W\tilde{Q}(\mathbf{y}_{[u]} - K(\mathbf{x}_{[u]})). \end{aligned} \tag{24}$$

Stages in the iteration of (24) are shown in figure 4. It too converges to $\mathbf{x}_a$, and it has the advantage that intermediate stages in the iteration have weights structured more like B, so intermediate $\mathbf{x}_{[u]}$ are more realistic, and the analysis errors stay less than the background errors.

The iterations shown in figures 3 & 4 started from $\mathbf{x}_{[0]}=\mathbf{x}_b$. The weights W are such that initially the increments are not in modes where $\mathbf{x}_b$ is accurate, so $\mathbf{x}$ will stay reasonably realistic. Thus it is common to omit the background term, and terminate the iteration before final convergence. For model-space and observation-space respectively, (12) and (24) become

$$\mathbf{x}_{[u+1]} = \mathbf{x}_{[u]} + QW(\mathbf{y}_o - K(\mathbf{x}_{[u]})). \tag{25}$$

$$\mathbf{x}_{[u+1]} = \mathbf{x}_{[u]} + W\tilde{Q}(\mathbf{y}_o - K(\mathbf{x}_{[u]})). \tag{26}$$

Stages in the iteration of these are shown in figures 5 & 6; particularly in the observation-space normalization version (26), the omission of the background term does not initially have a large effect, although it does affect the state after many iterations. The omission of the background term causes the iterations to converge towards a state which fits the observations exactly, i.e. at the position of each observation, that observation has net weight one, and all the others net weight zero.

3 ANALYSIS USING FILTERS

In general, the covariance matrix inverses B^{-1} and $(O+F)^{-1}$ are not operators which map the model and observation spaces onto themselves; that would be dimensionally incorrect. For that reason, they do not have physically meaningful eigenmodes. Filters, on the other hand, do map the spaces onto themselves. We can define a physically meaningful, non-dimensional inner-product for the model space (for instance a non-dimensional total energy) $\mathbf{x}^T P_x \mathbf{x}$, where P_x is a diagonal matrix of grid-box areas and scaling factors. We can hence define a non-dimensional transformed model space of

$$\hat{\mathbf{x}} = P_x^{1/2} \mathbf{x}. \tag{27}$$

The background error covariance in this space is given by

$$\hat{B} = \langle (\hat{\mathbf{x}}_b - \hat{\mathbf{x}}_t)(\hat{\mathbf{x}}_b - \hat{\mathbf{x}}_t)^T \rangle. \tag{28}$$

$\hat{B}$ does map the space onto itself; it has meaningful eigenvalues and eigenvectors, and we can decompose it (Lanczos 1957)

$$\hat{B} = U \Lambda U^T \tag{29}$$

$$\hat{B}^{-1} = U \Lambda^{-1} U^T \tag{30}$$

$$U U^T = I \tag{31}$$

where Λ is a diagonal matrix of the eigenvalues of $\hat{B}$, and the columns of U are the eigenvectors. Multiplying by $\hat{B}$ ($\hat{B}^{-1}$) is thus equivalent to projecting onto the eigenvectors, multiplying each by λ (λ^{-1}), and projecting back to model space. For covariance functions such as (16) which are specified as a function of distance only, the eigenvalues of $\hat{B}$ will be a function of the scale represented by the eigenvector, so $\hat{B}$ acts as a scale selective filter. The equivalent filter (R) in normal model space can be constructed by transforming to $\hat{\mathbf{x}}$ space, using $\hat{B}$ and transforming back:

$$R = P_x^{-1/2} \hat{B} P_x^{1/2}. \tag{32}$$

From (2) (27) and (28) we get

$$\hat{B} = P_x^{1/2} B P_x^{1/2}, \tag{33}$$

so

$$B = R P_x^{-1}. \tag{34}$$

This difference between R and B is important for models which do not have a uniform grid; R maps the model space onto itself, and can be thought of as a filtering operation, while B maps the dual of the model space onto the model space.

We can similarly represent (O+F) by a operator S which maps the observation space onto itself, and an inner product matrix for the observation space P_y:

$$(O+F) = S\,P_y^{-1}. \tag{35}$$

We need to use the inverse filter:

$$(O+F)^{-1} = P_y S^{-1}. \tag{36}$$

For the examples in this section we assume independent observations, without any error correlations, so $P_y=I$ is a suitable inner product, and S is a diagonal matrix of observation errors. Correlated observation errors are considered in section 4.

Using (10) (34) and (36) we have

$$W = RP_x^{-1}K^T P_y S^{-1}. \tag{37}$$

Now the adjoint K^* of K is defined in terms of the inner products in model and observation space:

$$(K^*y)^T P_x x = y^T P_y (Kx), \tag{38}$$

giving

$$K^* = P_x^{-1}K^T P_y. \tag{39}$$

So using filter operators, rather than covariances, leads to an equation using the adjoint of K, rather than its transpose:

$$W = RK^* S^{-1}. \tag{40}$$

To apply this, instead of calculating and multiplying by W, we operate on the the observational increments successively by S^{-1}, P_y, K^T, P_x^{-1}, and the filter R. To implement R, we could, in our simple example, calculate it exactly. However the aim of this work is to derive practical methods, so we use instead a recursive filter with response similar to the covariance function (16). Note that, as normally formulated, a low-pass filter leaves a constant field unaltered; that is not the case for R, which multiplies a constant field by the area average of the covariance function. We therefore include this factor in our recursive filter R. Full details are given in appendix A. In order to derive diagonal approximation to the normalizations Q or $\tilde{Q}$, as defined in (11) and (22), we apply the same idea as in (20), i.e. we operate on a large scale (spatially constant) field with the operators that make up W, and find the diagonal matrix that restores the original field. This can be done without calculating the individual weights w_{kl} used in (21). For our simple example, we are assuming the observations are similar, with no error correlation, so

$P_y S^{-1}$ is a diagonal matrix of $1/\epsilon_o^2$. Correlated observational errors are dealt with in section 4. In grid-point space, Q is the diagonal matrix such that

$$\mathbf{m}_x = Q\ (I+RP_x^{-1}K^T P_y S^{-1}K)\mathbf{m}_x, \tag{41}$$

where $\mathbf{m}_x$ is a constant model field. In observation space, $\tilde{Q}$ is the diagonal matrix such that

$$\mathbf{m}_y = \tilde{Q}\ (I+KRP_x^{-1}K^T P_y S^{-1})\mathbf{m}_y, \tag{42}$$

where $\mathbf{m}_y$ is a constant field of observations.

Figures 7-10 show the recursive filter applied to the four iteration schemes (12) (24) (25) and (26). They are very similar to results from the correlation function schemes shown in figures 2-5.

4 CORRELATED OBSERVATIONAL ERRORS

The examples thus far have assumed that O and F (and hence $P_y S^{-1}$), are diagonal matrices. In fact the instrumental error O is likely to be correlated if different observations use the same instrument, or processing method, which has any form of systematic error or bias. The representativeness error F contains contributions from the scales in the true atmosphere sampled by the observations but not resolved by the model. As the real atmospheric spectrum is not white, this leads to a correlation. Unfortunately, we have very little precise knowledge of actual observational error correlations. Schlatter and Branstator (1979) presented some evidence that satellite temperature profiles were correlated, but other empirical data is scarce. We therefore use physical arguments to derive a simple model for observational error covariances. We also have some justification for approximating the matrix inverse in the optimal iterative method.

Miscalibration of satellite instruments will cause very large scale correlations. Data processing methods are often tuned to "average" meteorological conditions; deviations from the average can cause correlated errors with the same scale as atmospheric correlations - larger than normal background error correlations. Thus there are grounds for expecting some observing systems to have a correlated component to the observational error. Yet each individual measurement has uncertainties which may not be correlated. The error of representativeness, although correlated on scales smaller than resolved by the model, has a significant uncorrelated component for normal observational spacings. We model observational errors for a typical satellite instrument by

$$(O+F) = E + C, \qquad (43)$$

where E ($=\varepsilon_u^2 I$) is the uncorrelated error covariance, and C is a correlated error covariance which can be modelled by a low-pass filter. In this case S will be a low pass filter, and S^{-1} will amplify smaller scales. The arguments in section 2 that justified (21) as an approximation for Q that guarantees convergence for all scales are no longer valid. Figure 11 shows analyses for the case where the observation errors are correlated with the same scale as the background errors, and $\varepsilon_u^2=0$. The scale dependence of S^{-1} and R cancel, so all scales converge rapidly in the iterative methods. If the observation error correlation scale is larger than that of the background, then Q may have to be increased to ensure convergence. Table 1 shows the inverse of the observational error covariance for the nine observations in our example, if 75% of the observational error variance is correlated, with twice the scale of the background error correlation. The column labelled sum shows the contribution factors to the normalization Q, if we calculate it as suggested in section 2 from the result of filtering a large-scale field. The numbers are small (even negative) so the normalization factors derived in this way are large, and iterations do not converge. Figure 12 shows an analysis obtained with the normalization factors reduced by a factor of three; this is sufficient to ensure convergence.

Accurate inverse of observational error covariance									Sum	ob coords	
8.6	-3.1	-.4	-3.1	-.9	.2	-.4	.2	.4	1.5	.6	.6
-3.1	10.2	-3.1	-1.0	-2.0	-1.0	.2	.0	.2	.5	0	.6
-.4	-3.1	8.6	.2	-.9	-3.1	.4	.2	-.4	1.5	-.6	.6
-3.1	-1.0	.2	10.2	-2.0	.0	-3.1	-1.0	.2	.5	.6	0
-.9	-2.0	-.9	-2.0	11.4	-2.0	-.9	-2.0	-.9	-.6	0	0
.2	-1.0	-3.1	.0	-2.0	10.2	.2	-1.0	-3.1	.5	-.6	0
-.4	.2	.4	-3.1	-.9	.2	8.6	-3.1	-.4	1.5	.6	-.6
.2	.0	.2	-1.0	-2.0	-1.0	-3.1	10.2	-3.1	.5	0	-.6
.4	.2	-.4	.2	-.9	-3.1	-.4	-3.1	8.6	1.5	-.6	-.6

Table 1. Inverse of observational error covariance matrix, for observations with a correlated error 0.1875(1+r/2)exp(-r/2) and an uncorrelated error 0.0625.

To evaluate the analyses shown in figures 11 and 12, the matrix S was actually calculated and inverted. The matrix inverse was applied to the observational increments ($y_o - K(x_{[u]})$) and the results were weighted or filtered as in the methods for uncorrelated observations. In practice, we want to avoid the cost of a matrix inversion. We need to approximate the inverse $(\varepsilon_u^2 I+C)^{-1}$. Let us consider the effect of $(\varepsilon_u^2 I+C)$ on large and small scales. C will multiply a large scale field by a factor β, equal to the correlated error variance times the area integral of the correlation function (see A12), and remove small

scales. $\varepsilon_u^2 I$ will multiply all scales by ε_u^2. Thus the inverse must divide large scales by $(\varepsilon_u^2+\beta)$, and small scales by ε_u^2. The filter $\beta^{-1}C$ will (approximately) pass only the large scales, and $(I-\beta^{-1}C)$ will (approximately) pass only small scales. So our approximate inverse could be given by

$$(\varepsilon_u^2 I+C)^{-1} \cong \left(\frac{1}{\varepsilon_u^2+\beta} \beta^{-1}C + \frac{1}{\varepsilon_u^2} (I-\beta^{-1}C) \right). \qquad (44)$$

This avoids any matrix inversion, and if the observations are in a regular pattern (e.g. from a scanning satellite instrument), multiplication by it can be implemented by a filter. If the observations are not sufficiently regular, then the filtering can be achieved directly as a convolution of the correlated part of the observational error covariance with the observation increments. β is the sum of the convolution weights. (44) would be accurate for a perfectly selective filter (e.g. if a constant miscalibration error gives a correlation with infinite scale). However for a more normal covariance filter, scales which are only partly damped by C are given too large weight. We can alleviate this by using a different scale separation filter $\hat{C}$ to construct the approximate inverse:

$$(\varepsilon_u^2 I+C)^{-1} \cong \left(\frac{1}{\varepsilon_u^2+\beta} \hat{\beta}^{-1}\hat{C} + \frac{1}{\varepsilon_u^2} (I-\hat{\beta}^{-1}\hat{C}) \right). \qquad (45)$$

Figure 13 shows the accurate spectral response of $(\varepsilon_u^2 I+C)^{-1}$ and the response of an approximate filter constructed using a $\hat{C}$ with a scale 0.4 times that of C. Also shown is the product of the response of this and the background error covariance. Note that this is larger for medium scales than for very large scales by a factor of about 2.4. The normalization factor necessary to give a convergent iteration for these critical scales can be derived from this. The eigenvalues of (I-Q(I+WK)) must have modulus less than one, so Q(I+WK) may multiply a mode by up to 2.0. The normalization factor calculated for a large scale mode is too large for a two-dimensional mode with the critical scale by about $2.4^2/2$ for convergence; it must be reduced by this factor. Table 2 shows the approximate inverse given by (45) for comparison with table 1, and figure 14 show the resulting analysis weights with the normalization factor reduced by a factor of three, as for figure 12. The difference from the optimal solution is small. Thus a filter approximation to the inverse, like (45), should be as good in practice as an accurate inverse of an observational error covariance which is not known with any certainty.

Approximate inverse of observational error covariance									Sum	ob coords	
11.3	-2.6	-.9	-2.6	-1.7	-.7	-.9	-.7	-.4	.8	.6	.6
-2.2	12.1	-2.2	-1.5	-2.2	-1.5	-.6	-.8	-.6	.7	0	.6
-.9	-2.6	11.3	-.7	-1.7	-2.6	-.4	-.7	-.9	.8	-.6	.6
-2.2	-1.5	-.6	12.1	-2.2	-.8	-2.2	-1.5	-.6	.7	.6	0
-1.2	-1.8	-1.2	-1.8	12.8	-1.8	-1.2	-1.8	-1.2	.7	0	0
-.6	-1.5	-2.2	-.8	-2.2	12.1	-.6	-1.5	-2.2	.7	-.6	0
-.9	-.7	-.4	-2.6	-1.7	-.7	11.3	-2.6	-.9	.8	.6	-.6
-.6	-.8	-.6	-1.5	-2.2	-1.5	-2.2	12.1	-2.2	.7	0	-.6
-.4	-.7	-.9	-.7	-1.7	-2.6	-.9	-2.6	11.3	.8	-.6	-.6

Table 2. Approximate inverse, given by (45), of observational error covariance matrix, for observations with a correlated error 0.1875(1+r/2)exp(-r/2) and an uncorrelated error 0.0625.

5 APPLICATIONS

There have been many practical applications of these methods, often developed in a more ad hoc way than in this paper. Bergthorsson and Doos (1955) proposed a normalization of the form (21), but in early versions did not iterate. Cressman (1959) described a successive correction scheme. Barnes (1973) studied the spectral properties of a similar method. These early methods all used normalization in model space, and did not include the background term. They limited the number of iterations, and to speed convergence for smaller scales, reduced the radius of influence in later iterations. Figures 15 and 16 show the effect of this; they show better results than figures 5 and 6, the equivalents with constant scale. The grid-point normalized version still shows excessive extrapolation away from the data region; in practical schemes the influence region is limited even on the first iteration. The observation normalized version gives very close to the optimal weights (figure 2b) after only four iterations, whereas without the reducing scale more than eight iterations are needed.

Bratseth (1986) described a scheme normalized in observation space, with the background term included using (24). He demonstrated that it converged to the optimal solution. For multivariate analysis, e.g. of height and wind, the "large-scale mode" argument used to derive the expression (21) for the normalization factor are not valid. Bratseth suggests

$$q_{kk} = \left(\sum_l |w_{kl}| + 1 \right)^{-1}, \qquad (46)$$

which guarantees convergence.

The Meteorological Office "Analysis Correction" data assimilation scheme (Lorenc *et al.* 1991) also uses observation space normalization; this was found

to give better results than the model space normalization used in an earlier version. The multiplication by the weight matrix W is the most expensive part of the this scheme. W has a column for each observation, and an entry for each grid-point with a non-negligible correlation with the observation. Thus as the density of both grid-points and observations increase, the cost of calculating W will increase as their product. On the other hand the cost of K^T is proportional to the number of observations only, and the cost of a recursive filter is proportional to the number of grid points only. Thus the current increases in both model resolution and observation density give us the incentive to consider filter methods.

A recursive filter successive correction method (Purser and McQuigg 1982) is used in the Meteorological Office mesoscale analysis system. It has been adapted and extended to three dimensions by Hayden and Purser (1988). Grid-point normalization is used, with ad hoc tuning to optimize convergence in a limited number of iterations.

Derber and Rosati (1989) used a Laplacian filter in an analysis of sea-surface temperatures. Rather than our normalization factor, they regarded the iteration as a variational descent algorithm, and used the conjugate gradient method. This approach requires additional calculation for each iteration, but should converge more rapidly. The normalization method described in section 2 does not give the optimal step-length except for the very largest scales. Stability for the iterative method is still maintained for step-lengths up to twice as large. Figure 17 shows iterations with an increased step-length; it can be seen by comparison with figure 3 that convergence is quicker. Descent algorithms normally determine the step-length by searching along the descent direction for the minimum of the penalty function; this cannot be done directly unless the term $(\mathbf{x}_b-\mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x})$ can be calculated. Parrish and Derber (1990) have developed an iterative spectral statistical analysis, which assumes that the background error covariance has a simple (diagonal) form in spectral normal mode space. They can thus implement multiplication by B by scaling the mode amplitudes, a filtering operation, and can evaluate $\mathbf{B}^{-1}$ in the penalty function.

The iterative methods reviewed in this paper can form a simple and effective algorithm for analysing observations, leaving the option of using available computer power to increase the resolution of the model, the density of observations used, and the speed of analysis. It is important, as modern prediction systems improve, that the background information they can provide

is weighted correctly, compared to the various types of observation, so the ability to relate the methods to one giving optimal weights is desirable. The understanding gained helps design practical schemes for processing the high resolution information coming from remote sensing.

ACKNOWLEDGEMENTS

This work benefited from the ideas and suggestions of R J Purser.

APPENDIX A PROPERTIES OF RECURSIVE FILTER

There is an extensive literature on filters, which has been reviewed by Raymond and Gardner (1991). Here we are content to describe the simplest recursive filter which is adequate for our needs: to match the filtering of convolution with a covariance function. We follow the notation of Purser and McQuigg (1982), and Hayden and Purser (1988). We consider a one-dimensional grid with spacing δx; A_i is the initial value at grid-point i, B_i is the value after filtering for i=1 to n, and C_i is the value after one pass of the filter in each direction. α is the filter coefficient. Our basic recursive filter is defined by:

$$\begin{aligned} B_i &= \alpha\, B_{i-1} + (1-\alpha)\, A_i \quad \text{for } i=1,n \\ C_i &= \alpha\, C_{i+1} + (1-\alpha)\, B_i \quad \text{for } i=n,1 \end{aligned} \tag{A1}$$

This is a first order recursive filter, applied in both directions to ensure zero phase change. Multi-pass filters are built up by repeated application of (A1). If all observations are inside the area being filtered, then the initial field being filtered is zero outside the boundary, and we can determine boundary conditions that reproduce the effect of extending the area to infinity. Hayden and Purser (1988) give formulae for such boundary conditions, becoming increasingly complicated for later passes. For the first pass we have $B_0=0$ and $C_{n+1}=(\alpha/(1+\alpha))B_n$.

To analyse this filter, it is easiest to consider the inverse filter, which is not recursive. Rearranging (A1) we get

$$A_i = C_i - \frac{\alpha}{(1-\alpha)^2}\left(C_{i-1} - 2C_i + C_{i+1}\right). \tag{A2}$$

Thence we can calculate the spectral response of the inverse filter, which is the reciprocal of the response one the recursive filter (S). For a wave of length $2\pi/k$:

$$S_{1pass}(k) = \frac{1}{1 + \frac{\alpha}{(1-\alpha)^2}\left(2\sin\left(\frac{k\,\delta x}{2}\right)\right)^2}. \qquad (A3)$$

For an N pass filter, this factor will be applied N times. If $k\,\delta x$ is small and α is small, then

$$S_{Npass}(k) \simeq 1 - \frac{\alpha}{(1-\alpha)^2} N\, k^2 \delta x^2. \qquad (A4)$$

We need to compare this filter with the effect of convolving a field with a covariance function. Often, covariance functions are assumed to be Gaussian shaped; in one dimension being

$$b(x_i, x_j) = \varepsilon_b^2 \exp(-(x_i - x_j)^2/2R^2). \qquad (A5)$$

The spectral response of this is

$$S_{Gaussian}(k) = \varepsilon_b^2 \sqrt{2\pi}\, R \exp(-R^2k^2/2). \qquad (A6)$$

Note that the response for a constant field (k=0) is not unity; the recursive filter was designed to conserve the mean value of a field and has S(0)=1. So if we want a filter that has the same low-pass (k small) properties as the Gaussian shaped covariance filter, we need to multiply the result of the recursive filter by a factor β:

$$\beta = \varepsilon_b^2 \sqrt{2\pi}\, R, \qquad (A7)$$

and choose α such that

$$\frac{\alpha}{(1-\alpha)^2} N\, \delta x^2 = \frac{R^2}{2}. \qquad (A8)$$

Setting $E = N\delta x^2/R^2$ we get:

$$\alpha = 1 + E - \sqrt{E(E+2)}. \qquad (A9)$$

Figure 18 shows the effect of operating on a delta function at x=0, with a Gaussian shaped filter, and with 1, 2 and 3 pass filters with coefficients defined as above. This gives the equivalent correlation functions for x=0 (defined in (A5) for the Gaussian). As N tends to ∞, the N pass filter tends to a Gaussian filter. Note however that our method of matching different filters at large scales does not ensure that ε_b^2 for the equivalent covariance functions is the same. This, and the smoothing of smaller scales, are as important as the low-pass properties. The Gaussian shaped covariance function has been criticized for having insufficient power in small scales; a simple function with more is the second order auto-regressive (SOAR)

$$b(x_i, x_j) = \varepsilon_b^2 (1 + |x_i - x_j|/s) \exp(-|x_i - x_j|/s). \qquad (A10)$$

This has spectral response

$$S_{SOAR}(k) = \varepsilon_b^2 \frac{4s}{(1+k^2s^2)^2}. \qquad (A11)$$

Note that it is impossible to match this to a Gaussian shaped covariance simultaneously for the variance ε_b^2 and for the low pass response; matching the low pass response requires ε_b^2 for the SOAR to be $(\pi/2)^{1/2}$ times that for the Gaussian. However, for k δx small (i.e. for grid-length<<scale), it is possible to match the SOAR to a recursive filter; the auto-regressive origin of the SOAR can be related to a recursive filter. (A11) is similar to (A3) for N=2. Thus it is possible to define the scaling factor

$$\beta = \varepsilon_b^2 \, 4s, \qquad (A12)$$

and calculate the filter factor α from (A9) with $E=N\,\delta x^2/4s^2$, giving a recursive filter which, for 2 passes, with k δx small, is equivalent to convolution with a SOAR covariance function.

In two dimensions, the SOAR function is defined in terms of the distance between the points, while the recursive filter applied in orthogonal directions is equivalent to a correlation which is the product of one dimensional SOARs. For a Gaussian shaped covariance function this would not matter, but for the SOAR the area integral gives a spectral response for the mean field (k=0):

$$S_{SOAR2D}(0) = \varepsilon_b^2 \, 6\pi \, s^2. \qquad (A13)$$

This has $6\pi s^2$ instead of $(4s)^2$; there is a larger contribution from large distances. Thus the one-dimensional equivalence with a recursive filter is not exactly preserved in two-dimensions. In the examples shown in the paper we use a two-pass, two-dimensional filter with

$$\beta = \varepsilon_b^2 \, (4s)^2. \qquad (A14)$$

This does not have exactly the same low-pass response as the SOAR with which it is compared, but it does imply the same background error variance ε_b^2.

REFERENCES

Balgovind,R., Dalcher,A., Ghil,M., and Kalnay,E. 1983
"A stochastic-dynamic model for the spectral structure of forecast errors". Mon.Wea.Rev. 111, 701-721.

Barnes, S.L. 1973
"Mesoscale objective analysis using weighted time-series observations". NOAA Tech Memorandum ERL NSSL-62.

Bergthorsson, P and Doos B.R. 1955
"Numerical weather map analysis". Tellus 7 329-340.

Bratseth, A.M. 1986
"Statistical interpolation by means of successive corrections". Tellus 38A 439-447.

Cressman, G.P. 1959
"An operational objective analysis scheme". Mon, Wea, Rev 87 367-374.

Derber, John C. and Rosati, Anthony. 1989
"A global oceanic data assimilation system". J. Phys. Oceanogr., 19, 1334-1347

Hayden, Christopher M. and Purser,R James, 1988
"Three-dimensional recursive filter analysis of meteorological fields". Preprints, AMS 8th Conference on Numerical Weather Prediction, Baltimore, Md. February 1988.

Lanczos,C. 1957
"Applied Analysis". Prentice Hall.

Lorenc, A.C. 1986
"Analysis methods for numerical weather prediction." Quart. J. Roy. Met. Soc. 112, 1177-1194.

Lorenc,A.C., Bell,R.S., and Macpherson,B. 1991
"The Meteorological Office "Analysis Correction" data assimilation scheme". Quart. J. Roy. Met. Soc. 117, 59-89

Parrish,D. and Derber,J. 1990
"Direct analysis of model normal modes using optimal interpolation". Preprints, WMO Int. Sym. on Assimilation of Observations in Meteorology and Oceanography. Clermont-Ferrand France.

Purser,R.J. and McQuigg,R. 1982
"A successive correction analysis scheme using recursive numerical filters". MetO11 Tech.Note.154.

Raymond,William,H. and Gardner, Arthur. 1991
"A review of recursive and implicit filters". Mon.Wea.Rev. 119, 477-495.

Schlatter, T.W., and Branstator, G.W. 1979
"Errors in Nimbus 6 temperature profiles and their spatial correlation". Mon. Wea. Rev. 107, 1402-1413.

Thiebaux,H.J., Mitchell,H.L. and Shantz,D.W. 1986
"Horizontal structure of hemispheric forecast error correlations of geopotential and temperature". Mon.Wea.Rev., 114, 1048-1066.

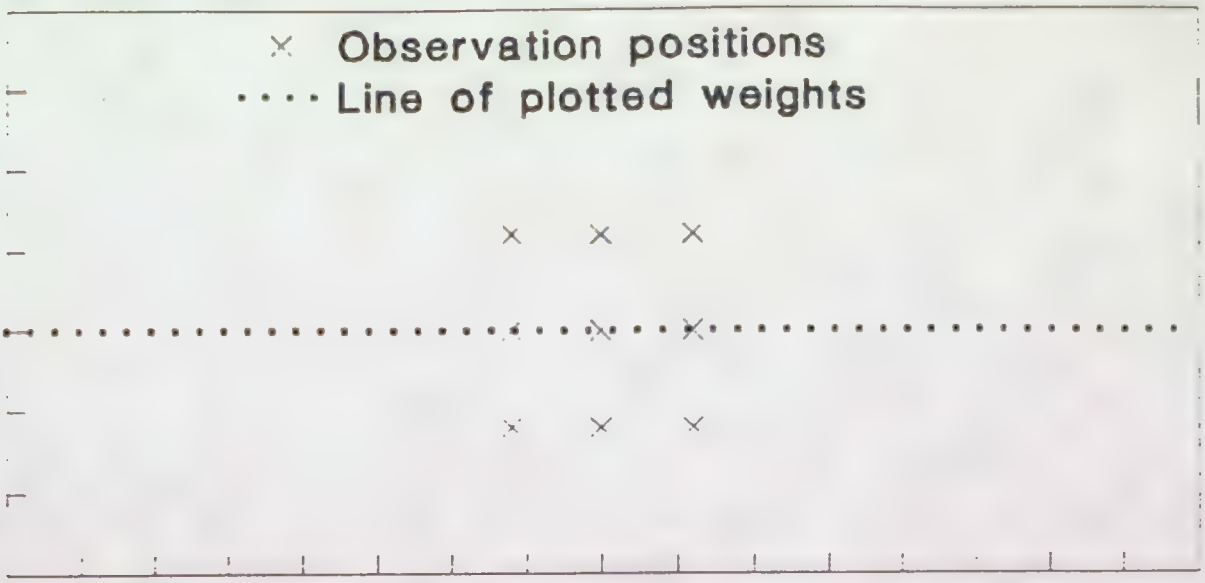
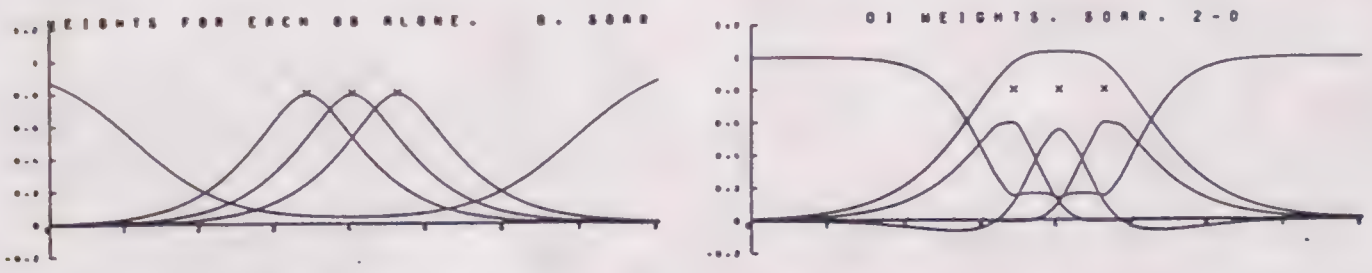


Figure 1. Position of observations used in examples, and line alone which weights are plotted. The ticks are separated by the distance s used in the SOAR formula (1+r/s)exp(-r/s).



(a) (b)

Figure 2. (a) Weights that the three observations on the line shown in figure 1 should have if each was alone for an optimal analysis. The x's show the position of the observations, and are repeated on subsequent figures for reference. Also shown is q as given by (21). (b) Weights given to three observations in an optimal analysis, the sum of weights for all nine observations, and the analysis error.

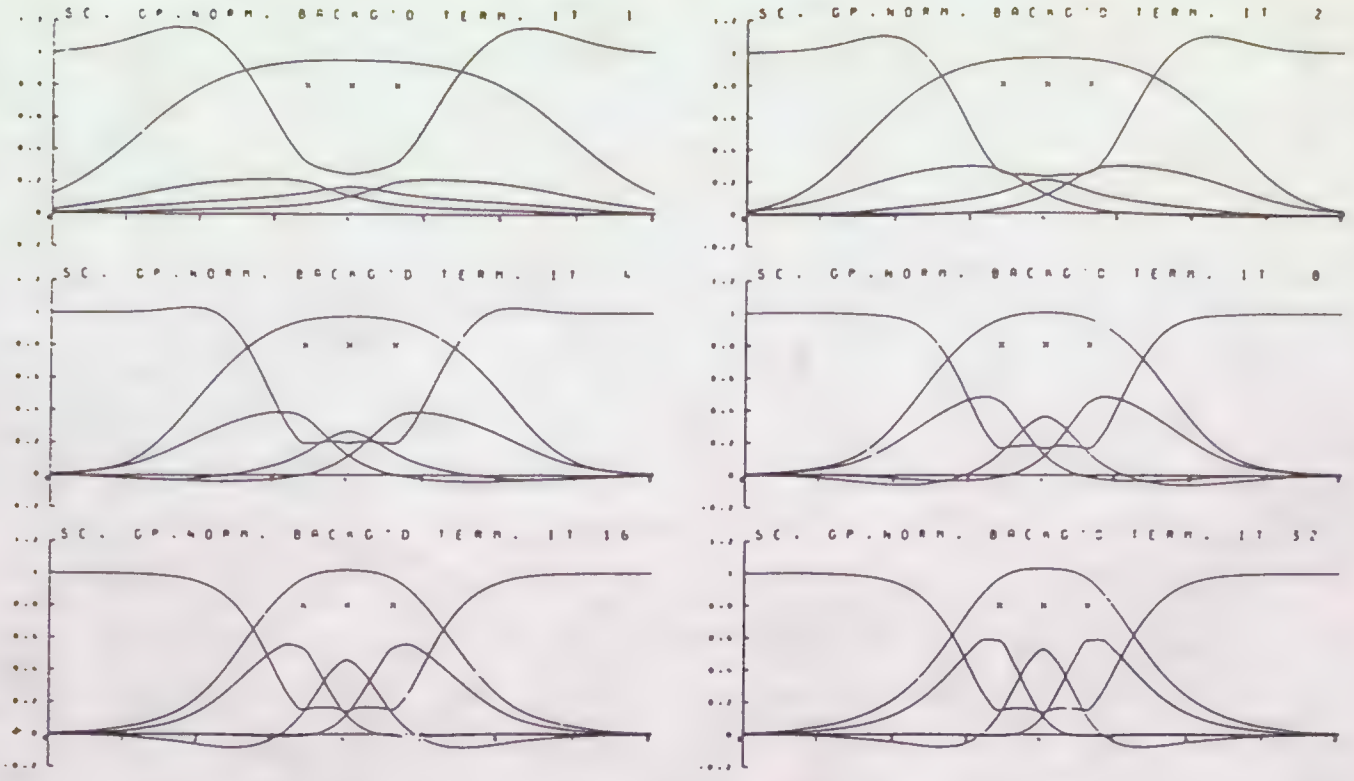


Figure 3. Weights given to three observations, the sum of weights for all nine observations, and the analysis error, after 1, 2, 4, 8, 16, and 32 iterations of the grid point normalized equation (12), using the approximate Q (21).

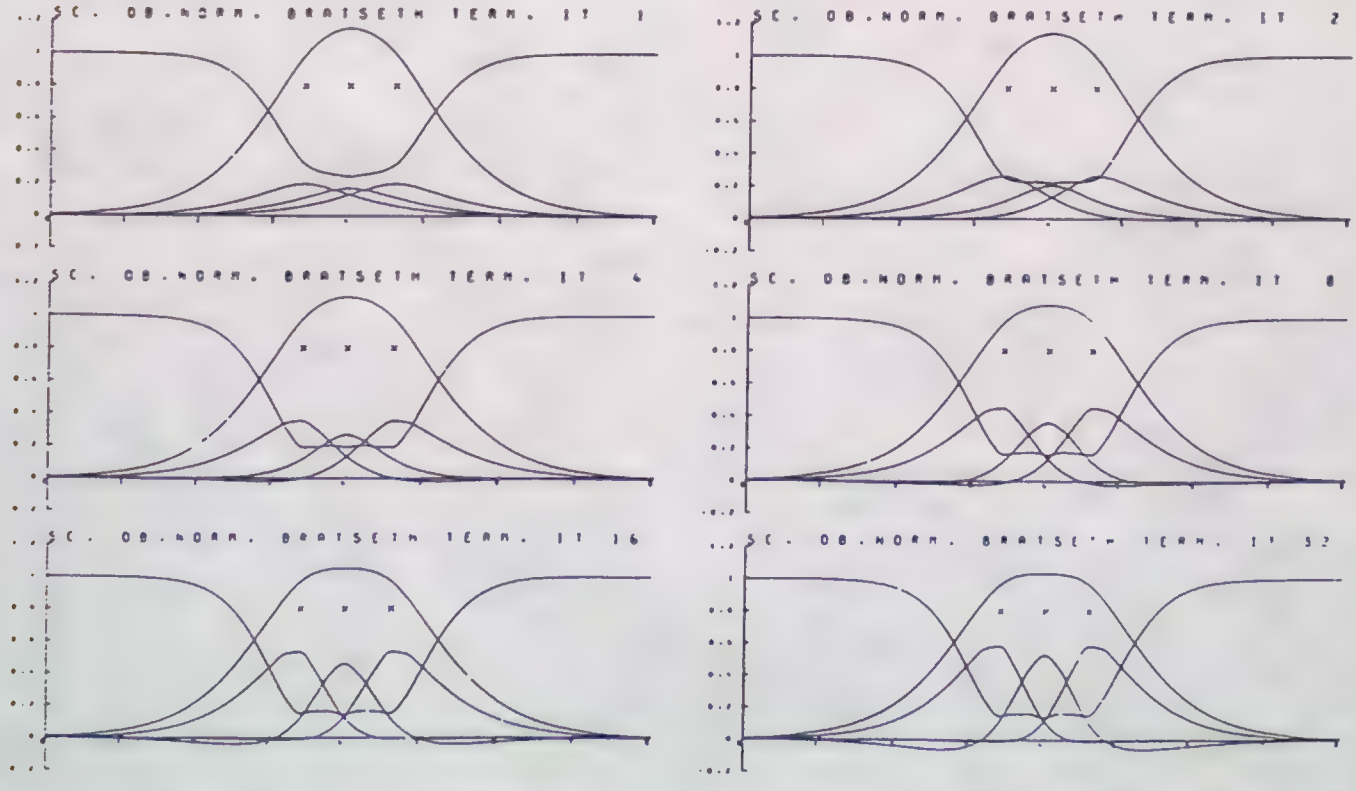


Figure 4. As figure 3 for the observation space normalized iteration (24).

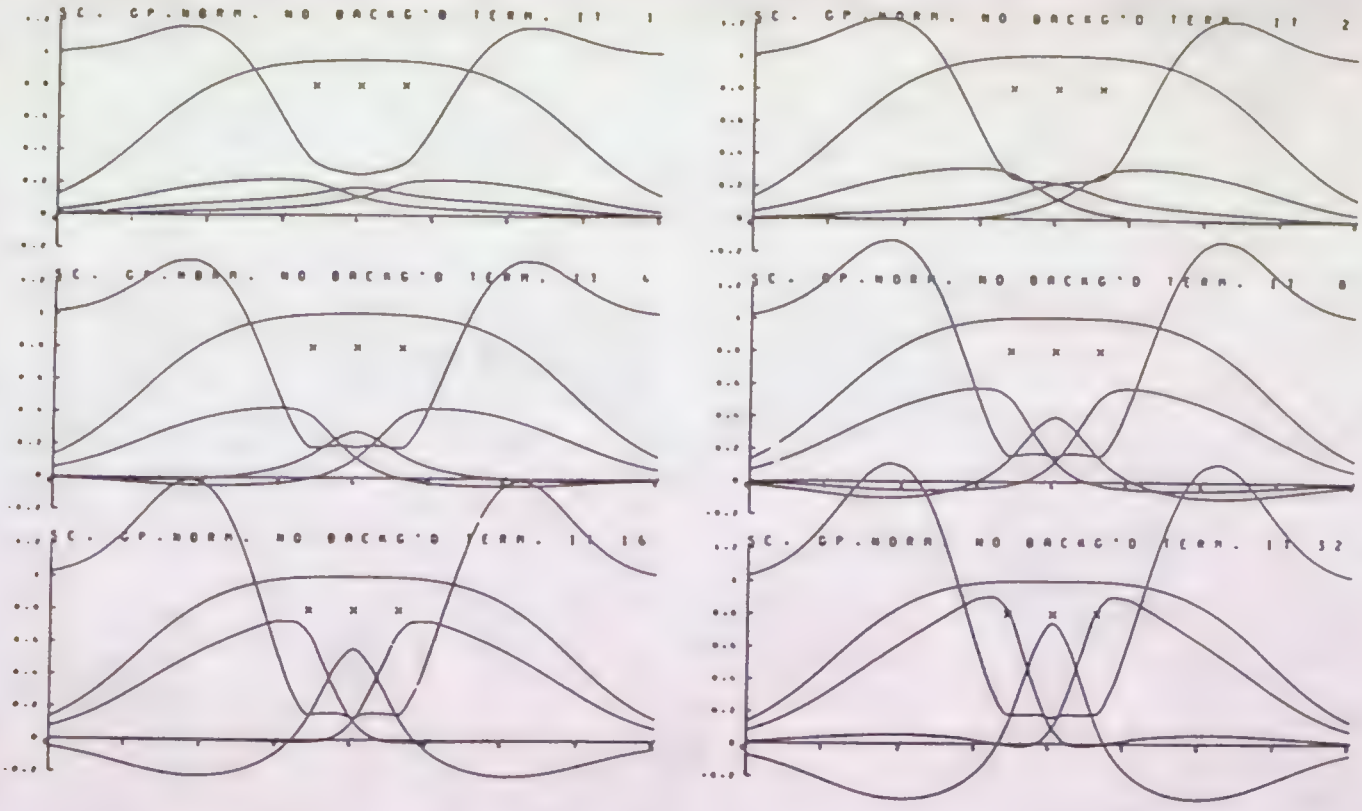


Figure 5. As figure 3 for the grid-point normalized equation omitting the background term (25).

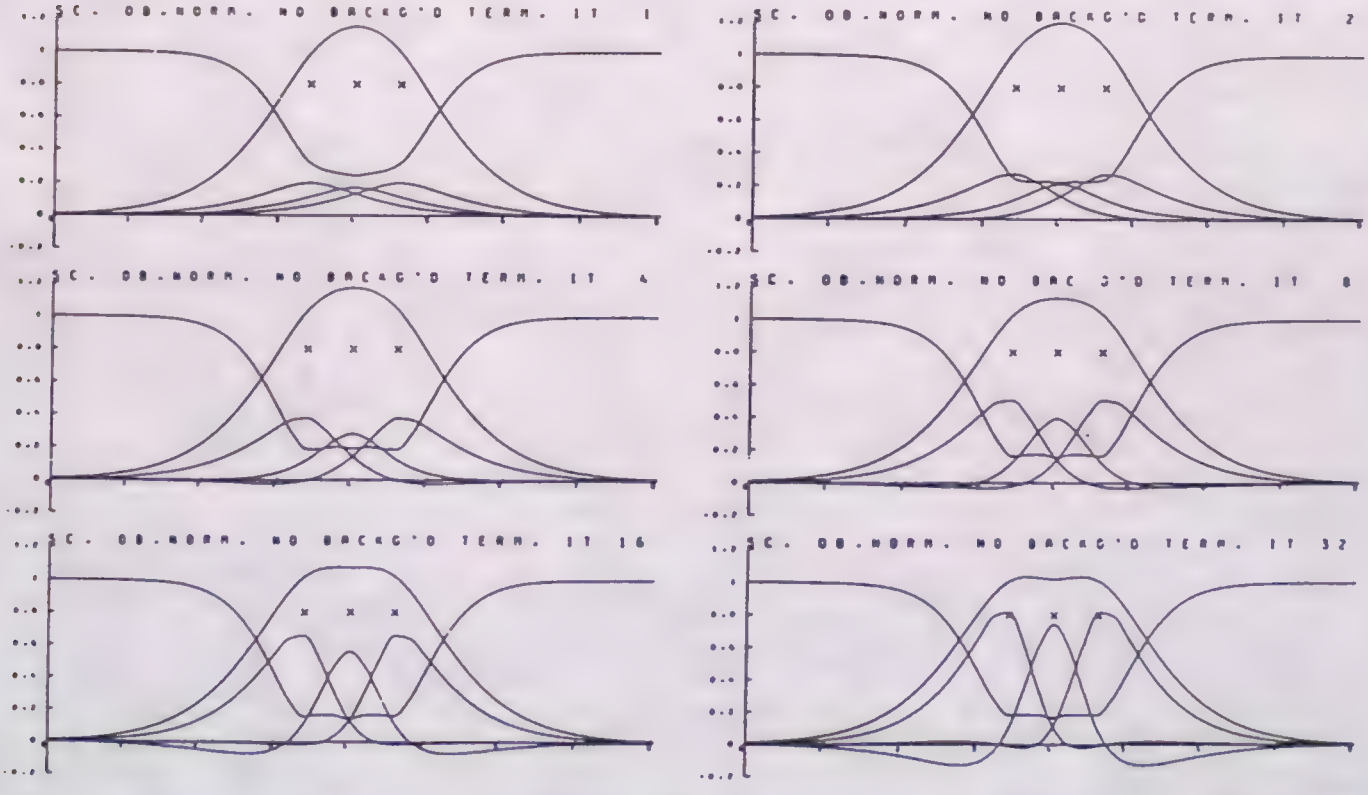


Figure 6. As figure 3 for the observation normalized equation omitting the background term (26).

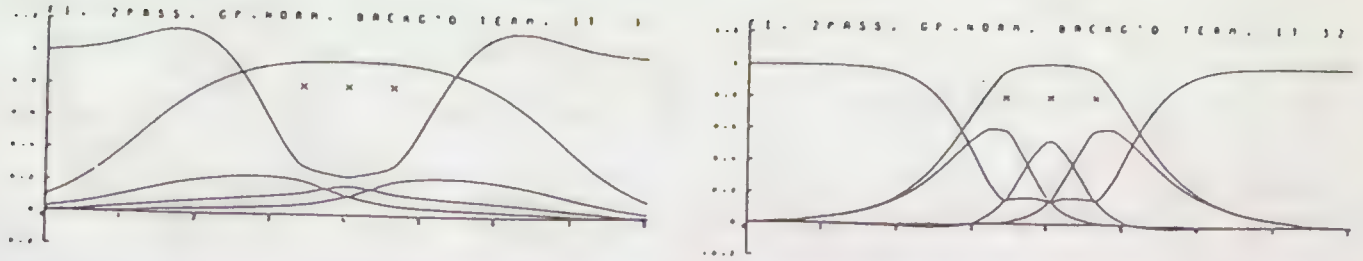


Figure 7. As figure 3 using the recursive filter, with grid-point normalization (iterations 1 and 32).

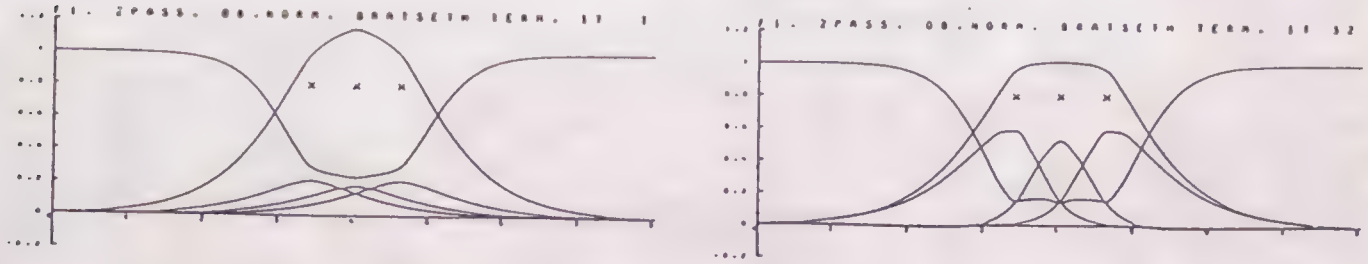


Figure 8. As figure 4 using the recursive filter, with observation normalization (iterations 1 and 32).

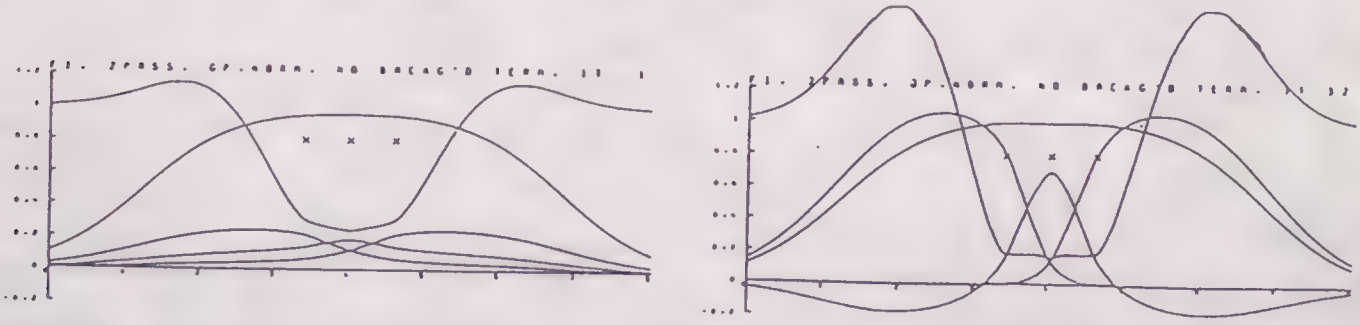


Figure 9. As figure 5 using the recursive filter, with grid-point normalization and no background term (iterations 1 and 32).

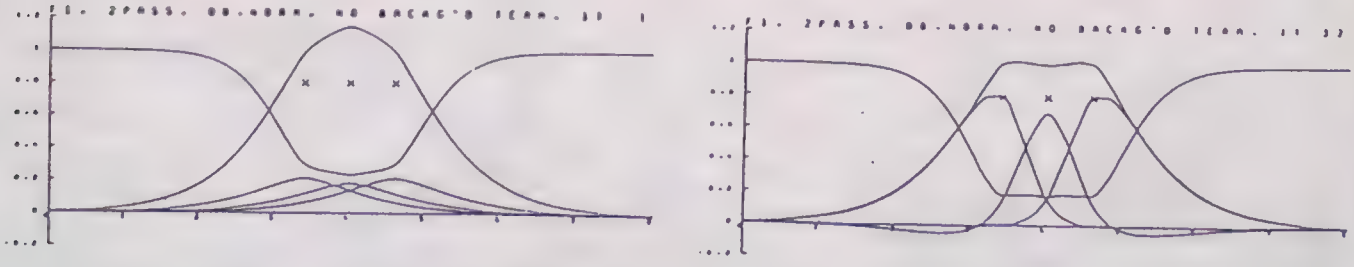


Figure 10. As figure 6 using the recursive filter, with observation normalization and no background term (iterations 1 and 32).

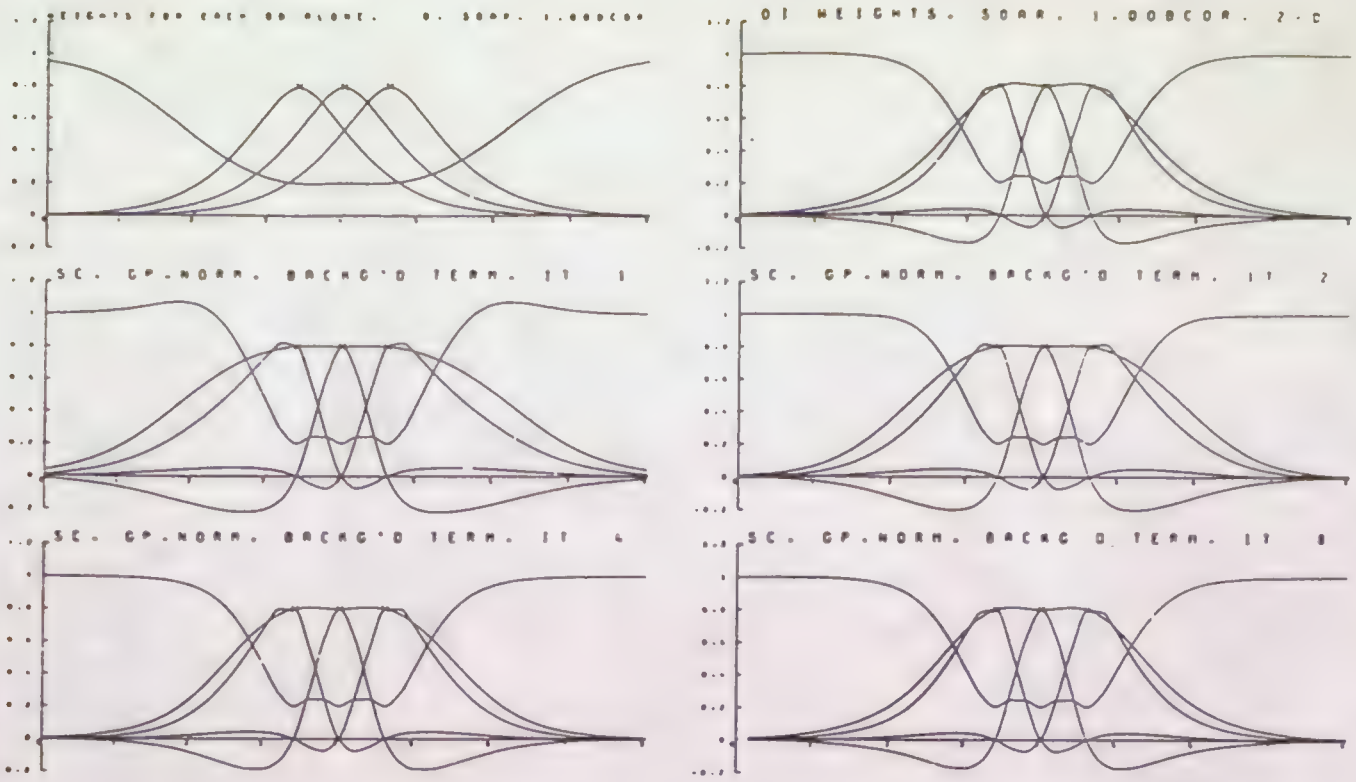


Figure 11. (top) As figure 2, (middle and bottom) as figure 3 iterations 1 2 4 and 8, for observations with the same error variance, but correlated like the background error. Grid-point normalization was used.

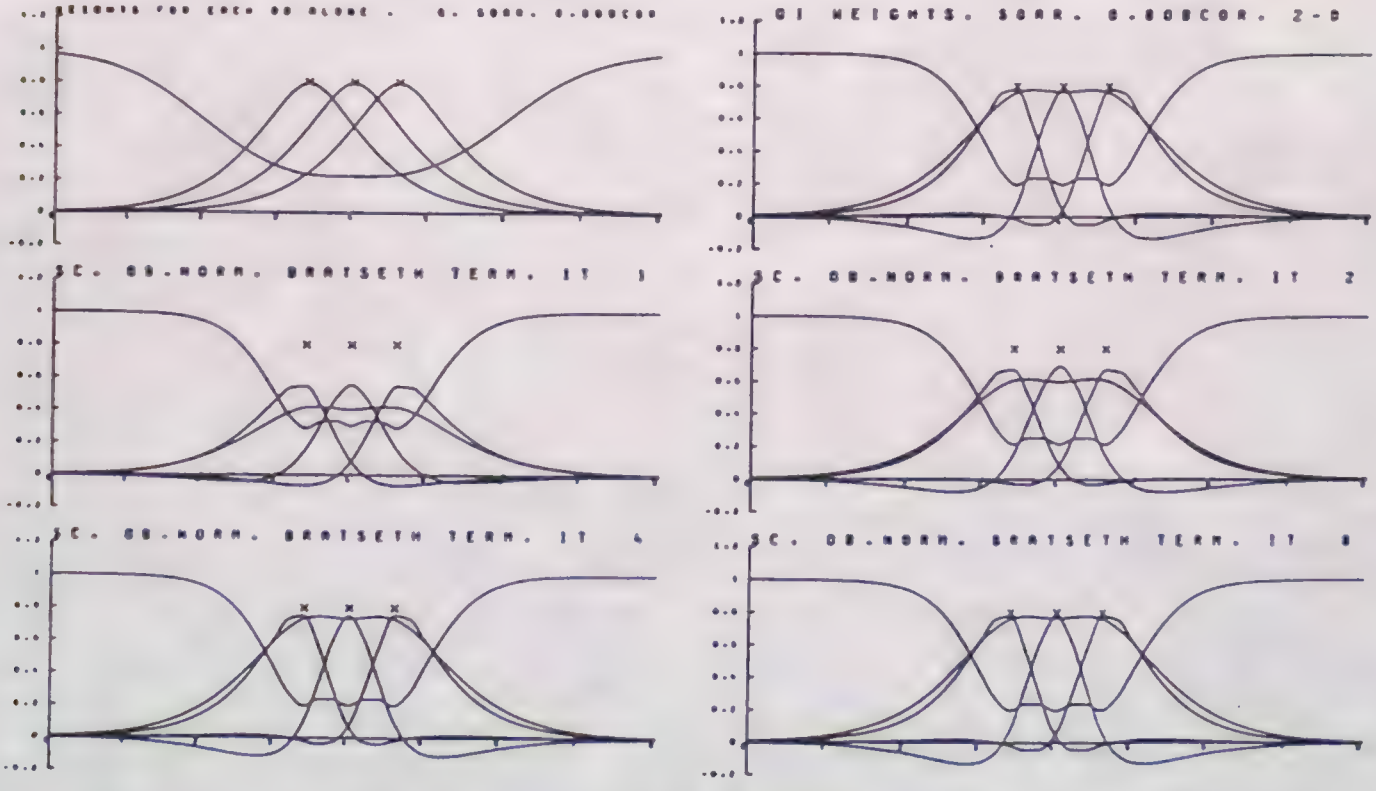


Figure 12. (top) As figure 2, (middle and bottom) as figure 4 for observations with the same error variance, 75% of which is correlated with a scale twice that of the background error. Observation normalization was used.

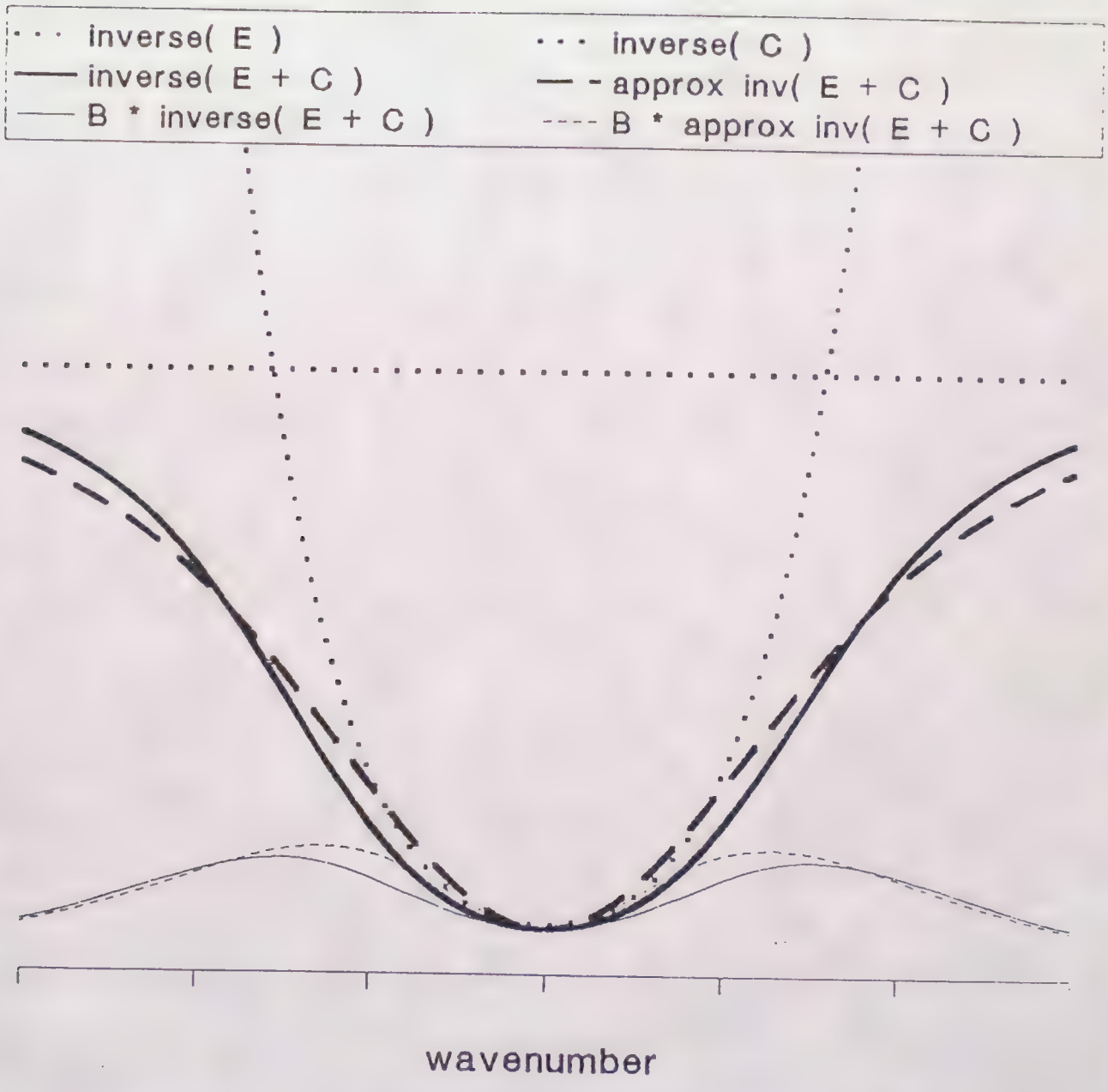


Figure 13. (thick lines) spectral response of inverse covariance filter, and its approximation (45). In the centre, for small wavenumber (i.e. large scales) the response matches that of the inverse of the correlated part of the covariance (shown dotted); at large wavenumbers it matches that of the inverse of the uncorrelated part (horizontal dotted line). Also shown (thin) are the products of of these filter responses with that of the background error covariance filter B.

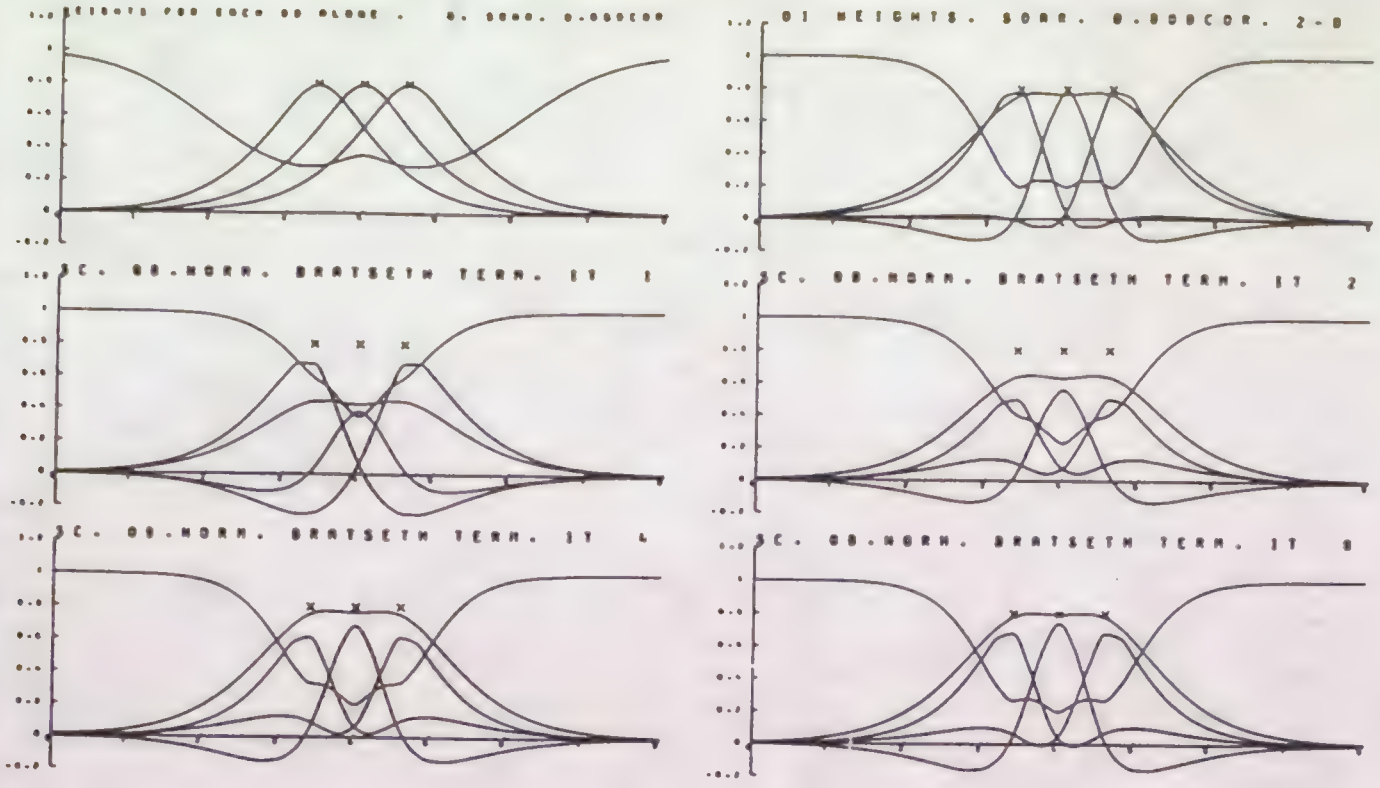


Figure 14. As figure 12 for an analysis performed using the approximate inverse observational error covariance filter (45).

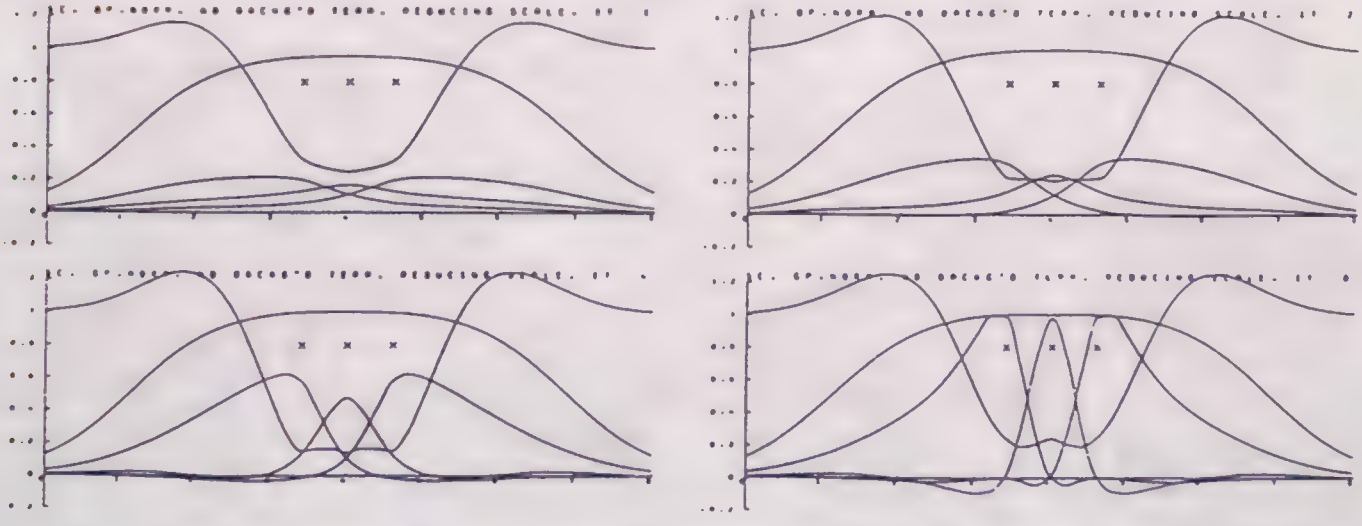


Figure 15. As figure 5 (grid-point normalization, no background term) with the correlation scale reduced by a factor of 0.8 after each iteration.

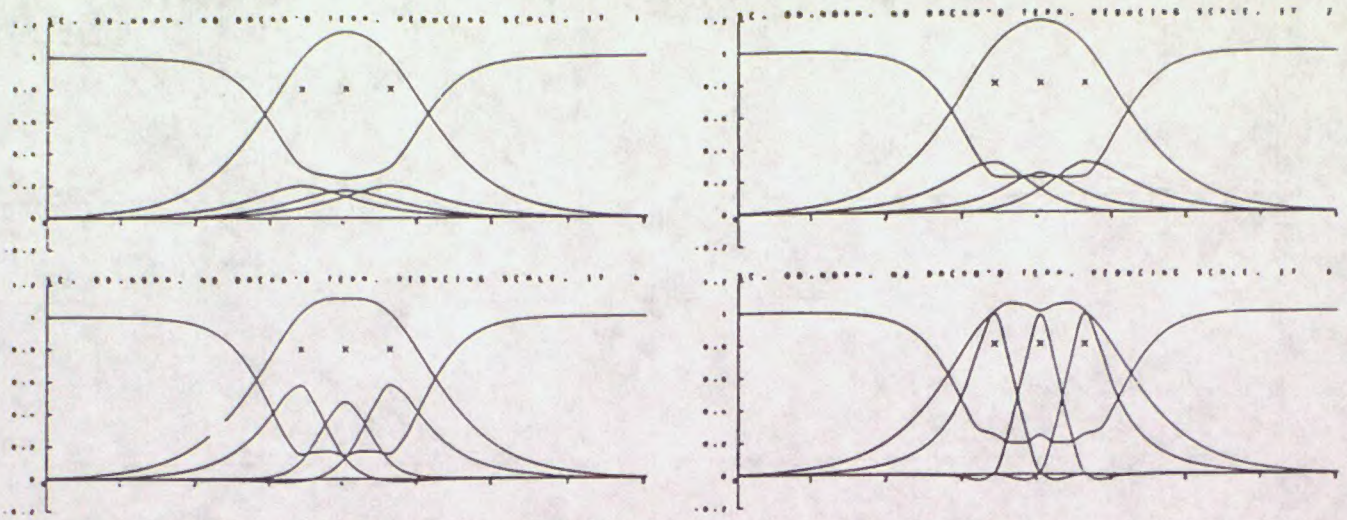


Figure 16. As figure 6 (observation normalization, no background term) with the correlation scale reduced by a factor of 0.8 after each iteration.

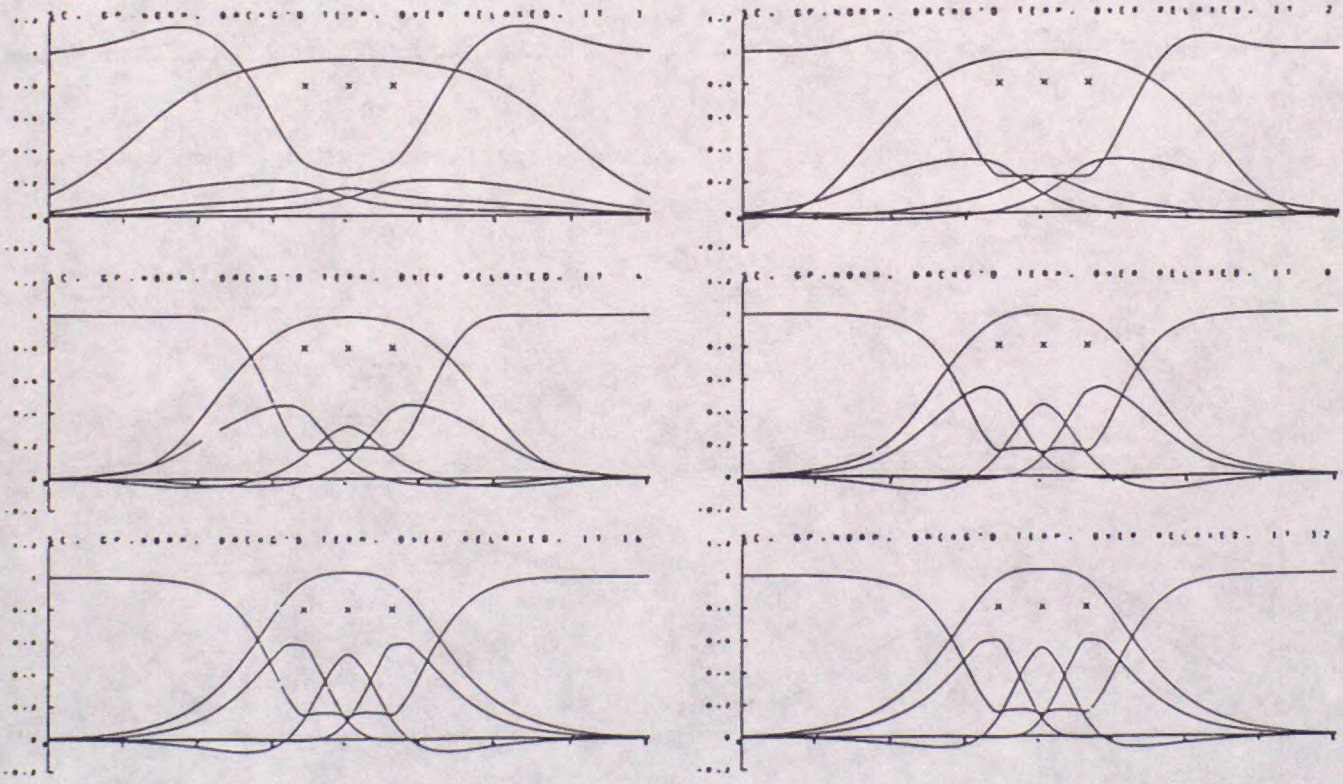


Figure 17. As figure 3 with Q increased by a factor (2 - 1/iteration no.).

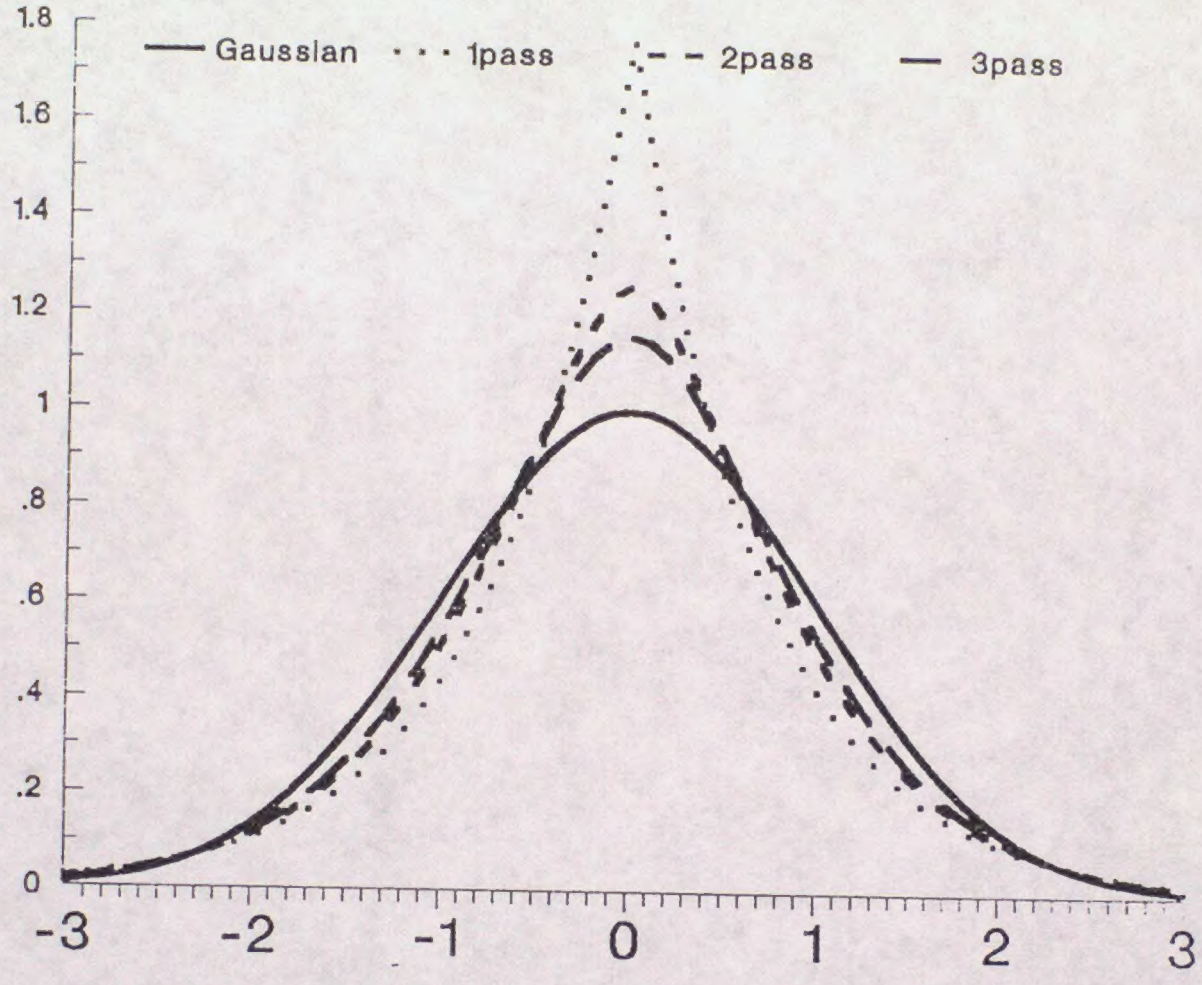


Figure 18. Gaussian shaped covariance function ($\exp(-x^2/2)$), and the equivalent covariances obtained by filtering a delta function at x=0 on a one-dimensional grid, with one, two and three pass recursive filters. Tick marks indicate the grid spacing of 0.1 units. The curve for a SOAR covariance ($(\pi/2)^{1/2}(1+2x)\exp(-2x)$) is indistinguishable from the 2 pass recursive filter curve. All the filters have the same response for long waves.

CURRENT SHORT - RANGE FORECASTING RESEARCH SCIENTIFIC NOTES (JULY 1991)

The Short- Range Forecasting Research Branch (formerly Met O 11) Scientific Notes which contain information of current use are listed below. The complete set of Scientific Notes is available from the National Meteorological Library on loan, if required.

3. **A geometric model of balanced, axisymmetric flows with embedded penetrative convection.**
G.J. Shutts, M. Booth and J. Norbury
February 1988

4. **Implicit finite difference methods for computing discontinuous atmospheric flows.**
M.J.P. Cullen
March 1988

5. **Variational aspects of semi-geostrophic theory.**
R.J. Purser
August 1988

6. **On the incorporation of atmospheric boundary layer effects into a balanced model.**
M.J.P. Cullen
July 1988

7 **Implicit finite difference methods for modelling discontinuous atmospheric flows.**
M.J.P. Cullen
June 1988

8. **An analytical model of the growth of a frontal intrusion.**
M.W. Holt and G.J. Shutts
November 1988

9. **Planetary semi-geostrophic equations derived from Hamilton's principle.**
G.J. Shutts
July 1988

10. **Semi-geostrophic moist frontogenesis in a Lagrangian model.**
M.W. Holt
September 1988

11. **Generalised Lagrangian solutions for atmospheric and ocean flows.**
M.J.P. Cullen, J. Norbury and R.J. Purser
November 1988

12. **Properties of the Lagrangian semi-geostrophic equations.**
M.J.P. Cullen and R.J. Purser
January 1989

13. **A simple two phase precipitation scheme for use in numerical weather prediction models.**
B.W. Golding
June 1989

CURRENT SHORT - RANGE FORECASTING RESEARCH SCIENTIFIC NOTES (JULY 1991) CONTINUED

14. **A test of a semi-implicit integration technique for a fully compressible non-hydrostatic model.**
M.J.P. Cullen
September 1989

15. **Dynamical aspects of the October storm 1987 : A study of a successful Fine-mesh simulation**
G.J. Shutts
September 1989

16. **A conservative split-explicit integration scheme suitable for forecast and climate models.**
M.J.P. Cullen and T. Davies
October 1989

17. **The dynamical structure of some North Atlantic depressions simulated by the fine-mesh model.**
G.J.Shutts
January 1990

18. **An Idealised Simulation of the Indian Monsoon using Primitive Equation and Quasi - Equilibrium models.**
M.H. Mawson and M.J.P. Cullen
August 1990

19. **Time - Integration Schemes for meteorological simulations involving oscillatory and dissipative modes.**
R.J. Purser
July 1990

20. **Mesoscale Model Experimental forecasts of the Haar of North East Scotland.**
S.P. Ballard, B.W.Golding, R.N.B. Smith
August 1990

21. **A conservative split-explicit integration scheme with fourth order horizontal advection.**
M.J.P. Cullen and T. Davies
October 1990

22. **A Geometric Model of the Nonlinear Equilibration of Two-Dimensional Eady Waves.**
M.J.P. Cullen and I. Roulstone
February 1991

23. **The Non-linear Evolution of a slowly growing wave on a laterally sheared baroclinic flow.**
M.J. Bell
April 1991

24. **Iterative Analysis Using Covariance Functions and Filters**
Andrew Lorenc
July 1991